Slide 1:

- Good afternoon, everyone, my name is Tuan Nguyen and today I will present my project on rainfall prediction using LSTM model

Slide2:

- The content of my presentation will include the introduction of the topic, followed by data and preprocessing steps, after that is the Model and training details, experimental results and finally the discussion

INTRODUCTION

Slide 4:

- I think everyone knows what rainfall is. But for a formal definition, it is the condensed water vapor that falls by gravity
- Rainfall can affect many different human activities such as: agriculture, transportation, industrial and energy sectors.
- Too few or too much rainfall can cause drought or flooding, both can have adverse effects on peoples live in general
- Therefore, it is quite important that we need an accurate and consistent way to predict when these rainfall events happen, to better prepare to mitigate or take advantage of its effects.

Slide 5:

- People have been trying to predict rainfall for thousands of years using many different methods. However, now adays, instead of simply praying to the god and ask for revelation, researchers have developed 2 main ways of rainfall predictions:
- 1 is based on physical simulation of the earth's atmospheric system, where the process of rainfall is physically simulated using data from ocean current and cloud movement on a global scale. 2 main examples of this type of model are NOAA's GFS system, or ECMWF's IFS system. While these models are highly reliable, they are also expensive to operate, and require extra work for small scale prediction.
- The other method is data driven models using historical record data to predict future rainfall, without relying on physical simulation. ML/DL models fall into this kind of method, and many have used different types of models on different types of data to extract information on future rainfall events. However, unlike simulation models, statistical based models do not adapt quickly to newer data, or sudden changes in weather characteristics.
- Both types of models work well in short term predictions, which are future prediction up to 14 days ahead. However, beyond this range, their prediction contains a lot of uncertainties and inaccuracy.
- The main problem with longer range prediction comes from the many influencing factors related to rainfall such as ocean current, ocean oscillation phrases, different weather cycles and effects, which are hard to measure accurately.
- This difficulty can be represented by the phrase "butterfly effect", where a butterfly flapping its wing could lead to a chain of events that cause the path of a tornado to change or prevent it entirely.
- This leads to weather prediction being largely non-deterministic, as everyone would often hear weather reporter saying that there are 80% chance of rainfall tomorrow, not that there will be rainfall tomorrow
- This uncertainty in prediction is a large problem in long term prediction, where errors in each timestep are magnified to produce results that do not correlate with real life situations.
- There have been a number of papers that address this problem, some of which suggested that instead of performing daily predictions, predictions should be performed on weekly or monthly data instead.
- The result of these models would often say that it would rain by a certain amount in a month, but not when the event will occur, or that it would be a single rainfall event, or multiple events within a month. As such, the temporal information of rainfall events was lost in the model result.

Slide 6

- In order to address this issue, this project proposes a model that would perform daily prediction for long term period, up to 30 days ahead. This model would utilize LSTM architecture with the aim to capture long term dependency between different weather measurements data and the rainfall, and to predict the future rainfall amount. The model will use meteorological measurement of 30 – 90 days prior to the prediction period as input, 2 variants of the model was created, the Daily iterative model, which will output 1 day ahead prediction, and Single Prediction model, where the model will predict 30 future days at once.

DATASET

Slide 8

- Moving onto the dataset, the dataset chosen for this project is the "rain in Australia" dataset from Kaggle. It has over 145k observations of 23 columns. This data contain daily measurement from nearly 50 stations across the country and contain information on different meteorological measurements, such as temperature, humidity, windspeed, evaporation, etc

Slide 9:

- Here is an example of the raw data, taken from the Australian Bureau of meteorology.

Slide 10:

- It is common for weather data to have missing values, and for this data set, the percentage of missing value for Evaporation, Sunshine hours, and Cloud are high, at almost 40-50%.
- The main reason for this huge amount of missing data is because some weather stations do not have the instrument to record specific measurements, or their gauge was not operational. Either way, these missing data are often specific to a weather station.
- In order to ensure the model performance would not drop during training, is to remove data from stations with high percentage of missing values

Slide 11:

- This slide shows the heat map of percentage of missing values, and based on this heatmap, all stations with above 30% of missing values in a column would be dropped from the dataset, apart from major cities with more than 1million inhabitants.

Slide 12:

- The final list contains 21 stations from the initial 49. Here is the full list of 21 selected stations, and their location within Australia.
- The final dataset contains over 64k observations of 23 columns.

Slide 13:

- Besides many missing data, this data set has zero inflated problem, as up to 2/3 of the total observations have rainfall amount equal to 0, which could cause potential problems to the model performance.

Slide 14:

- Here is a chart of rainfall frequency for each day of the year. On average everyday has around 20% chance of rainfall event occurring, and there doesn't seem to be any distinct pattern at first glance.

Slide 15:

- Moving onto the data processing: for missing values, the dataset is divided into smaller data for each station, and missing values will be filled based on the mode and mean of each stations, as each stations can have different mode/mean for each column, and using a single mode/mean for the entire dataset can lead to larger information loss.
- After missing values are replaced, the data are rescaled using MaxMin and Standard scaler from sklearn, while humidity was converted from % to numbers between 0-1

Slide 16:

- Since the model will use 30-90 days prior as input, the data was split into sequences of different length depending on the input length. This figure shows how the sequences are created by using sliding window for 60 days input and 30 days output.
- Next, the data set is split into train/val/test sets with ratio of 80% train, 15% validation and 5% test and loaded into Torch Data Loader.

MODEL

Slide 18:

- This slide shows the overall model architecture, The sequences data contain input weather data is sent to the LSTM cell, while the one-hot encoded predicting month would be sent to a fully connected layer, where the result of this layer would be added on top of the hidden layer output of LSTM cell.
- This result is then passed to 2 more linear layers to get the final output. The output size of the last linear layer is 1 for daily iterative variant, and 30 for single prediction variant.

Slide 19:

- Since this is a regression problem, the 2 most common loss functions for this kind of problem are the MSE and MAE. MSE loss will obtain guide the model to predict closer to the mean while MAE loss will pull the model prediction closer to the median.
- For rainfall data, both would have advantages and disadvantages. For advantages, MSE would be better at predicting outliers, but MAE would give better general prediction, as outliers are assigned the same weight as all other samples. For disadvantages, neither prediction would be good enough, as the predicting the mean will mean that everyday would have a little bit of rainfall, while predicting median would result in all predictions will be zero, as the median of the dataset is 0.
- A possible solution is to use a combination of MSE and MAE, which is known as Huber Loss. Simply put, a threshold of delta is assigned, and if the absolute difference between predicted and real value is larger than delta, MAE loss is used, and if its smaller, MSE loss is used.
- Huber loss is similar to the midway point between the 2 loss, combining both advantages of MSE and MAE loss function.

Slide 20:

- After the loss function was selected, the model was trained using default Adam optimizer parameters, with different input length and batch size tested, and the weight with best validation loss was saved as the best model weight.
- An interesting point to note was that the shuffling of data loader can also affect model performance, with some cases resulting in all 0 prediction models.

EXPERIMENT RESULT

Slide 22:

- Starting with the Daily Iterative Variant, on the left is a table containing results from grid training with different batch sizes and input sequence for daily iterative model. Generally , the longer the input sequence, the better the model performs. The 2 best configurations are models using 90 days input, with batch sizes of 64 and 256. The model with batch size 256 was selected since it has a balanced result between MAE and MSE score.
- On the right is a scatter plot showing the predictions on the y axis, and real values in the x axis. There is a slightly higher concentration on points above the line compared to below it, indicating that the model slightly over predicts compared to actual values.

Slide 23:

- Here is a plot showing the actual prediction on test set. We can see that the model has managed to pick up the general trend of data but is having difficulty correctly guessing the right amount.

Slide 24:

- Here is a residual plot of the prediction, and the model do make a lot of mistakes on both ends.

Slide 25:

- Moving onto the Single prediction variant, due to time and resource constraints, the model was trained using 60days sequence input and batch size of 128. Compared to the daily iterative, it has worse performance, as this model is trying to predict 30 future days at once.
- On the right we can see there is a large concentration of points below the line, indicating that the model is severely under-predict the magnitude of rainfall events.

Slide 26:

- Here is the prediction for the test set, we can see the model constantly underpredicts on the test set.

Slide 27:

- To better gauge the model's performance, this table contains test scores for the best trained model, a model that only predicts 0, and a model with randomized weight.
- For Daily Iterative variant, the trained model performed better compared to the other 2. However, for the Single prediction variant, the trained model has better MSE but worse MAE. A possible explanation is that the model sacrifice MAE to make more non-zero prediction to better match outliers, which can lower MSE but increase MAE.

DISCUSSION

Slide 28

- Moving onto the discussion, the first thing I would like to talk about is the effect of loss function on model training. Besides Huber loss, MAE and MSE loss were also tested during model training. From the prediction results on the right, we can see how loss function can affect the model behavior and can lead to model being stuck in some local optima

Slide 29:

- The difference in behavior between the 2 variants is also interesting, as the only difference between these variants is the output size of the last layer. From the result on the right, which is the prediction of these models for March 2023 of Sydney, we can see how the Daily Iterative model tends to overshoot the amount and under predict occurrence, while the other variant do the exact opposite.
- It is an interesting behavior and would require further testing to clarify.

Slide 30:

- Moving onto the model limitations, currently, both variants of the model rely only on the last hidden stage of LSTM to perform prediction, which is not the best design, specially in case of Single prediction model, where this last stage was used to predict 30 days ahead. This is one of the common limitations of vanilla LSTM model, along with model ability to utilize previous hidden stages and its ability to retain information when moving through the long sequence of 90 days.
- These limitations might jog your memory, as this is something Professor had mentioned in class, with the LSTM translator model

Slide 31:

- To address these limitations, in future works, the most obvious solution would be to incorporate attention mechanism to the model, either by using a modified LSTM mode, or using Transformer model as basis to create a more efficient rainfall predictor.

THANK YOU, THE END